Content-preserving Tone Adjustment for Image Enhancement

Simone Bianco
Viale Sarca 336, 20126 Milano, Italy
University of Milano-Bicocca
simone.bianco@unimib.it

Claudio Cusano
via Ferrata 1, 27100 Pavia, Italy
University of Pavia
claudio.cusano@unipv.it

Flavio Piccoli
Viale Sarca 336, 20126 Milano, Italy
University of Milano-Bicocca
flavio.piccoli@unimib.it

Raimondo Schettini
Viale Sarca 336, 20126 Milano, Italy
University of Milano-Bicocca
schettini@unimib.it

Input image

Test image enhanced by the proposed system

Abstract

We propose a novel method based on Convolutional Neural Networks for content-preserving tone adjustment. The method is at the same time fast and accurate since we decouple the inference of the parameters and the color transform: the parameters are inferred from a downsampled version of the input image and the transformation is applied to the full resolution input. The method includes two steps of image enhancement: the first one is a global color transformation, while the second one is a local transformation. Experiments conducted on the DPED — DSLR Photo Enhancement Dataset, that has been used for the NTIRE19 Image Enhancement Challenge, and on the MIT-Adobe FiveK dataset, that is widely used for image enhancement, demonstrate the effectiveness of the proposed method.

1. Introduction

Tone adjustment is a non-linear operation for manipulating the color profile of an image to improve its visual quality. It is a particular case of image enhancement in which the colors are remapped with the aim of improving the dynamic range and of giving a more natural and pleasant appearance to the input image. Tone adjustment can be used to improve the pictures taken with portable devices to bring their quality to the level of better devices, such as DSLR cameras. With the huge increase of smartphones and the corresponding demand for better acquisition devices, a solution based on tone adjustment can dramatically change the quality perceived by the final user [1].

Image enhancement methods in the state of the art can be divided according to the type of transformation applied to the input image. A first group of methods uses neural networks to estimate implicitly the transformation at a pixel

level, making the enhancement a direct regression on the pixel values. Within this group, Isola et al. [11] propose a supervised method that makes use of adversarial training in combination with an L_1 loss over the RGB values to enhance the input image. While the L_1 and L_2 losses are known to promote blurriness [16], they favor the training convergence. In contrast, the adversarial loss is very effective in creating sharpen details. Zhu et al. [17] address the task of image enhancement as a semi-unsupervised task. The target of their work is to project the input image into the manifold of enhanced images, in an unpaired fashion. Yan et al. [15] introduce an image descriptor that accounts for the local semantics of the input image with the aim of allowing a more precise local enhancement. Ignatov et al. [8] propose to learn a translation function using a residual convolutional neural network to improve both color rendition and image sharpness. They use it to enhance images taken with smartphones to a quality level of a DSLR camera.

In contrast to these methods of direct estimation of the RGB values of the output images, there are the *conservative* methods, i.e. methods that preserve the content in the input image. These methods generally attempt to estimate a color transformation that is later applied to the values of the input image. Among these there are methods that estimate local or global transformations. In particular, Bianco et al. [2] estimate a patchwise color transformation which is later interpolated so that there is a color transform for each pixel of the input image. This approach gives the ability to make a spatially-varying enhancement without sacrificing the speed. Bilinear interpolation makes the enhancement transformations spatially smooth. Similarly, Gharbi et al. [6] adopt local color transformations on a subsampled version of the input image. Speed is greatly increased by limiting inference on a smaller version of the input image and applying the resulting transformations on the input image at full resolution. Bianco et al. [3] use a set of global color transformations to enhance the input image. By inferring a global instead of a local color transformation, the approach becomes more conservative by avoiding any spatial distortion. In addition, moving inference at a downscaled version of the input image makes the system suitable on mobile devices with low computational power.

To test new image enhancement methods, there are several benchmarks available to the researchers. They can be divided in two types of datasets: the *paired* and the *unpaired* benchmarks. The difference among these two types of datasets is that in the first case, each training sample is coupled with a per-pixel corresponding ground-truth, while in the second case there are two sets of pictures (input and enhanced) which may not have any relationship. Between these two groups, there is an *hybrid* group of benchmarks including samples which are semi-aligned, i.e. each input picture has a ground-truth which is not aligned per-pixel but depicts the same spot in the scene. An example of dataset belonging to this group is the DSLR Photo Enhancement Dataset (DPED) [8], that was created for the Challenge on Perceptual Image Enhancement on Smartphones [10] and for the NTIRE19 Image Ehancement Challenge [9], which includes crops of images acquired with an iPhone paired with images of the same spot acquired by a Canon DSLR camera (see subsection 3.1 for more details).

Getting inspiration from the work of Bianco et al. [3], we present a system able to enhance the input image. In particular, our contributions are:

- a fully end-to-end trainable system for image enhancement that can be trained either on paired and hybrid datasets;
- a fast, lightweight and scalable method for tone adjustment;
- a system designed to be conservative with respect to the content of the input image.

The rest of the paper is organized as follows: in Section 2 we will present the method, in Section 3 the experimental setup is described. In the same section, results are presented and compared with the state of the art. Finally, conclusions and further research directions are discussed in Section 4.

2. Proposed method

In this work, we propose a method for image enhancement that is conservative with respect to the content in the input image. Our method includes two sequential stages of enhancement. The first one is a *full color transformation* formed by a triplet of functions, one for each color channel, that combines all the color coordinates of the input pixel. Each triplet is composed by a set of three piecewise functions. Let (r_x, g_x, b_x) be the input pixel and (r_y, g_y, b_y) the corresponding output pixel. Then, the transformation is defined as:

$$c_y = c_x + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} \theta_{ijkc}\phi_i(r)\phi_j(g)\phi_k(b), \quad c \in \{r,g,b\}, \tag{1}$$

where $\vec{\theta} \in \mathbb{R}^{n\times n\times n\times 3}$ is the output of a first CNN (which acts as a parameter estimator) and

$$\phi_i(x) = \max\{0, 1-|(n-1)x-i+1|\}, \quad i \in \{1,2,\dots N\}, \tag{2}$$

is the basis piecewise function, assuming equispaced nodes in the range $[0,1]$.

The network that infers the parameters of these color transformations is described in Table 1. Inference is done on a scaled version of the input image having size 100×100 to speed up the computation. The network starts with a bilinear resampling to resize the input to 100×100 pixels.

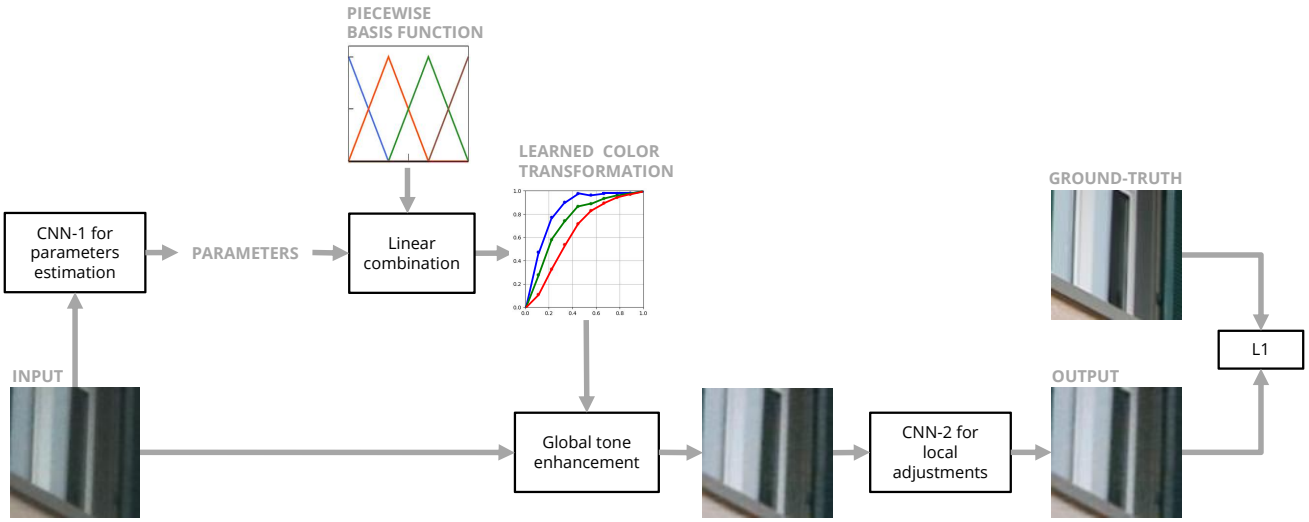


Figure 1. Pipeline of the proposed method. A neural network estimates the parameters of a global full color transformation composed by a triplet of piecewise functions, and applies this color transformation to the pixels of the input image. Finally, a second tiny network makes small local adjustments. In training time, the system is trained in a supervised manner with a L_1 loss.

Then there is a sequence of three convolutions, each one followed by a PReLU activation. Afterwards, the feature maps are averaged through the use of average pooling and processed by two linear layers interleaved by a ReLU activation. The result is an estimate of the $N^3 \times 3$ coefficients of the piecewise transformations, where N is the number of nodes.

The second enhancement step is necessary for training on hybrid datasets, where the input and the ground truth images are not perfectly aligned. This further step of enhancement consists of a spatial filter applied independently on each one of the three RGB channels. Let γ be the 2D filter. This operation is then:

$$c_y(w,h) = \sum_{i=-3}^{3} \sum_{j=-3}^{3} \gamma_{ij} c_x(w+i, h+j), \quad c \in \{r,g,b\}, \tag{3}$$

The filter is estimated through gradient descent and has a size of 7×7. To avoid artifacts on borders, the convolution is preceded by a reflection padding of size 3. Both enhancement stages are trained end-to-end in a single step using an L_1 loss. Figure 1 shows the complete pipeline including both enhancement stages.

3. Experiments and results

The proposed method has been evaluated on two different datasets, the DSLR Photo Enhancement Dataset (DPED) that has been created for the Challenge on Perceptual Image Enhancement on Smartphones [10] and the NTIRE19 Image Enhancement Challenge [9], and the MIT-Adobe FiveK dataset [4]. In this section, we will first present these two datasets and then the experimental setup used to assess the performance of the proposed system.

Table 1. Structure of the convolutional neural network used to estimate the coordinates of the nodes of the piecewise transformation. N denotes the number of nodes of the piecewise function and $H \times W$ the size of the input image.

Stage	Operation	Output size
Pre-processing	Input	$H \times W \times 3$
	Bilinear resampl.	$100 \times 100 \times 3$
Conv. Network	Conv. + PReLU	$49 \times 49 \times 16$
	Conv. + PReLU	$24 \times 24 \times 32$
	Conv. + PReLU	$11 \times 11 \times 64$
	Avg. Pooling	$1 \times 1 \times 64$
	Linear + ReLU	64
	Linear	$3N^3$
Post-processing	Color transf.	$H \times W \times 3$

Table 2. Architecture of the convolutional neural network used to compute local adjustments and to address the misalignment. The convolution is done channelwise with a 7×7 2D filter whose weights are learned during training.

Stage	Operation	Output size
Pre-processing	Input	$H \times W \times 3$
	Padding	$(H+3) \times (W+3) \times 3$
Spatial filter	2D Conv.	$H \times W \times 3$

3.1. Datasets

To assess the goodness of the proposed method, we demonstrate that is able to work on aligned as well as on semi-aligned datasets. The benchmarks used in this work are the DPED and the FiveK dataset.

DSLR Photo Enhancement Dataset (DPED) [8] The DPED dataset has been designed to enhance an image acquired with a portable device, in order to appear as if it was acquired with a better device. Each sample of the dataset is composed by two images depicting the same scene, acquired with a mobile phone and a Canon 70D DSLR, which are respectively the input and the ground truth. In the DPED dataset three different mobile phones were considered: an iPhone 3GS, a BlackBerry Passport and a Sony Xperia Z; however, in this work we only consider the iPhone images, since these are the ones onto which the NTIRE19 Image Enhancement Challenge [9] is based. This benchmark is provided in two formats: the original images at full resolution and a cropped version in which only crops of the original images are considered. The second variant is composed by crops having size 100×100 of semi-aligned pictures. Training, validation and test sets have respectively 160471, 4304 and 3057 samples. The fact that only small crops of pictures are given makes not possible to base the enhancement on semantic information. This is the benchmark that has been used to assess the quality of the methods participating to the Challenge on Perceptual Image Enhancement on Smartphones [10] and on the NTIRE19 Image Enhancement Challenge.

FiveK dataset [4]. The MIT-Adobe FiveK dataset is a dataset widely used for image enhancement. The dataset contains 5000 raw images enhanced by five different experts in the field of image enhancement, named expert A, B, C, D, and E. During the enhancement process, the experts were allowed to use only global color transformations. Since in the state of the art expert C is used for evaluation with this dataset, in this work we consider each sample composed by a raw image and the same image enhanced by expert C. The dataset is split in 4700 samples for training and two testsets of 50 and 250 samples each.

3.2. Error metrics

Following the guidelines of the NTIRE2019 Image Enhancement Challenge [9], the image enhancement algorithms are compared in terms of two commonly used metrics: the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity (SSIM) [14] index. For both metrics, the higher the better.

3.3. Parameters setting

We implemented the proposed system in python 3.7 and the Pytorch package [13] at the version 1.0.1.post2. The code of the implementation is available on GitHub[1]. We trained the method using a Nvidia® Titan Xp with 12GB of memory and 3840 CUDA cores. Weights of the convolutions are initialized with the method described in He et al. [7]. The number of nodes N used for the piecewise functions is 10. We train the model using Adam optimizer with the two betas used for computing running averages of gradient and its square respectively of value 0.9, 0.999; a learning rate of $1e^{-4}$ and a weight decay of 0.0. The system takes about 5 hours to converge.

3.4. Results

Results on the DSLR Photo Enhancement Dataset (DPED) [8] validation set. The results in terms of PSNR and SSIM are reported in Table 3. Figure 3 reports some visual results of our method on this dataset and the results achieved by the methods composing the state of the art. From the results it is possible to notice that the proposed method outperforms common methods in the state of the art. In addition, the design of the proposed system is such that the number of parameters is really low (up to 3 orders of magnitude less than other methods). Another crucial aspect is the speed of the system. In fact, as it is possible to observe in Figure 3, this method, not only is the most accurate in terms of PSNR, but it is also the fastest one. This fact, in addition to its lightness, makes the system adequate and ready to be adopted on mobile devices. As already said before, the image misalignment present in this dataset is addressed by the second enhancement module, that is composed by the 2D filter: in Figure 2 (a) it is possible to observe the learned weights of the filter. From the weights it can be noticed how the learned filters acts as a sort of low-pass filter.

Results on the DSLR Photo Enhancement Dataset (DPED) [8] test set. The results in terms of PNSR and SSIM are reported in Table 5. In addition to the 3057 100×100 crops, the challenge organizers have also provided 10 fullsize images to be processed. Figure 7 shows an example of full-resolution image enhanced by our system. Note that even though the method has been trained on small crops, it can be applied with good results to a high resolution images as well.

Results on Adobe FiveK [4] test set. The results in terms of PSNR and SSIM are reported in Table 4. Also in the case of perfectly aligned images (as it occurs in the FiveK dataset), the system outperforms the state of the art on

[1] https://github.com/dros1986/content-preserving-tone-adjustment-for-image-enhancement

Table 3. Quantitative comparison with state-of-the-art methods for single-image enhancement on the DSLR Photo Enhancement Dataset (DPED) [8] validation set used in the NTIRE2019 Image Enhancement Challenge [9], sorted by increasing PSNR.

Method	PSNR	SSIM
Pix2Pix [11]	21.43	0.75
HDRnet [6]	22.09	0.79
Cycle Gan [17]	22.09	0.87
Unfiltering [2]	22.28	0.78
Parametric [3] (dct)	22.31	0.77
Parametric [3] (p. wise)	22.31	0.77
Parametric [3] (poly.)	22.34	0.77
Parametric [3] (rbf)	22.38	0.78
Content-preserving tone adj. (ours)	**22.63**	**0.80**

Table 4. Quantitative comparison with state-of-the-art methods for single-image enhancement on *Expert C* of the Adobe FiveK dataset [4], sorted by increasing PSNR.

Method	PSNR	SSIM
Cycle Gan [17]	19.39	0.78
Unfiltering [2]	21.67	0.88
HDRnet [6]	22.31	0.89
Parametric [3] (poly.)	22.62	0.89
Parametric [3] (dct)	22.75	0.89
Parametric [3] (rbf)	22.87	0.89
Parametric [3] (p. wise)	22.94	0.89
Pix2Pix [11]	23.06	0.86
Content-preserving tone adj. (ours)	**23.14**	**0.90**

the expert C. In this case, the second enhancement is not needed, because images are perfectly aligned. As it is possible to observe in Figure 2, the system is able to deactivate this module in training stage by learning the identity function. This versatility makes this system ready-to-go on any dataset of image enhancement, no matter if it is semi-aligned or perfectly aligned. In addition, this method is the second fastest on this dataset.

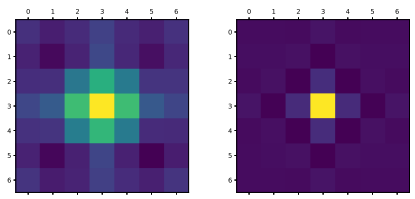
3.5. Processing time

One of the advantages of our method is that it allows to quickly process high-resolution images. The more complex operations are carried out on a low-resolution thumbnail and the time for the final application of the conservative color transformation is proportional to the number of pixels. Figure 5 shows the processing speed in frames per second (FPS) as a function of the size of the input image. For small images (100×100 pixels), the system is able to process the input image at more than 1200 frames per second (0.8ms per image). For large images (3000×3000 pixels) it is still possible to have a frame rate of 23 FPS (0.0431s).

Table 5. Quantitative comparison with methods participating to the NTIRE2019 Image Enhancement Challenge [9] on the DSLR Photo Enhancement Dataset (DPED) [8] test set, sorted by increasing PSNR.

Method	PSNR	SSIM	MOS
MENet	18.40	0.76	2.25
ViPr	18.69	0.73	2.29
Dong et al. [5]⋆	19.27	0.90	-
Ignatov et al. [8]⋆	20.08	0.92	-
Johnson et al. [12]⋆	20.32	0.92	-
MiRL	20.97	0.74	-
IVL (ours)	21.37	0.72	2.4
BOE-IOT-AIBD	21.74	0.78	2.55
Geometry	21.75	0.78	2.41
HIT-UltraVision	21.92	0.78	2.39
HIT-Xlab	22.14	0.79	2.53
TTI	22.17	0.76	2.53
Mt.Stars	22.35	0.79	2.78
TeamInception	22.41	0.79	2.60
BMIPL_UNIST_DW	22.44	0.80	2.59
Rainbow	22.66	0.80	-

⋆ Results computed with a different protocol, taken from [8].



(a) (b)

Figure 2. Learned filters of local post-processing on (a) the DSLR Photo Enhancement Dataset (DPED) [8] validation set and (b) on the FiveK [4] dataset. When the input and the ground-truth are aligned (case (b)), the final filtering is basically an identity, while if input and ground-truth are not aligned, this filter takes care of the slight misalignemnt and enables the training (case (a)).

4. Conclusions

In this work, we presented a method for tone adjustment that is able to work on aligned as well as semi-aligned datasets without having to make any specific calibration of the system. This method outperforms the state of the art both on the DSLR Photo Enhancement Dataset (DPED) [8] validation set as well as on the FiveK dataset [4]. Furthermore, this system is really lightweight (it has up to 3 orders of magnitude parameters less than other methods). The lack of a decoding part makes also the system extremely fast and suitable to be used on mobile devices. This method is also

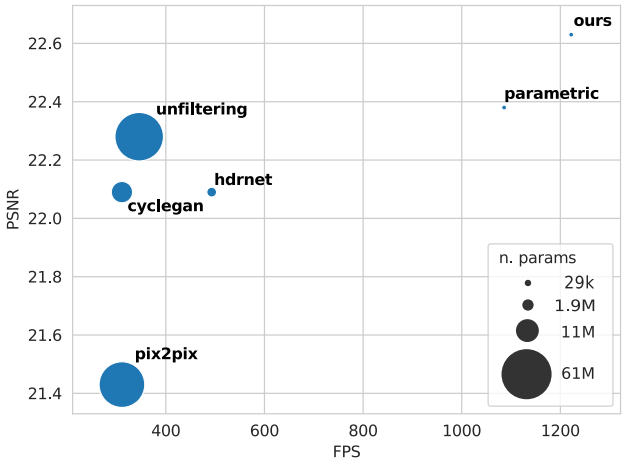


Figure 3. Sizes of the networks compared to achieved PSNR on the DSLR Photo Enhancement Dataset (DPED) [8] validation set. As it's possible to observe, our method is the one that achieves the highest PSNR while having the lowest number of parameters. This makes our system easier to learn and suitable for fast enhancements.

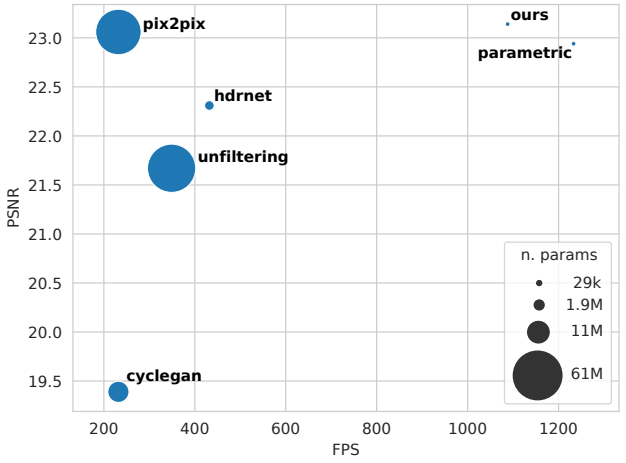


Figure 4. Sizes of the networks compared to achieved PSNR on the FiveK [4] dataset. Also in this case, our system outperforms other methods with a lot less parameters.

able to generalize well on never-seen images.

As a possible research direction, it would be interesting to develop a metric that estimates the result of a subjective study result. This would improve the perceived quality of the results.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

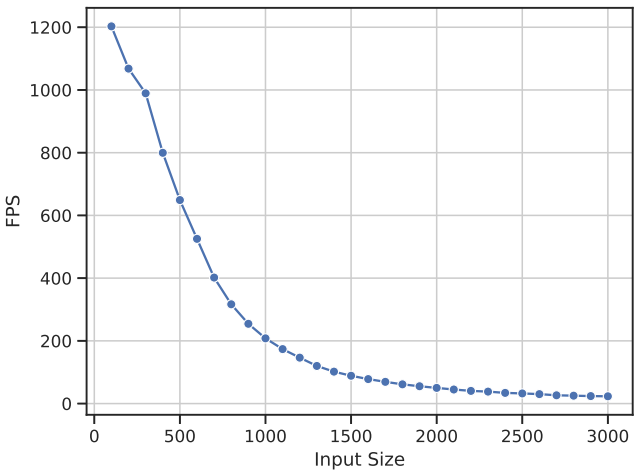


Figure 5. Processing speed of our system varying the size of the input image. Each point has been measured by processing 1000 batches composed by one image.

References

[1] Simone Bianco, Gianluigi Ciocca, Fabrizio Marini, and Raimondo Schettini. Image quality assessment by preprocessing and full reference model combination. In *Image Quality and System Performance VI*, volume 7242, page 72420O. International Society for Optics and Photonics, 2009. 1

[2] Simone Bianco, Claudio Cusano, Flavio Piccoli, and Raimondo Schettini. Artistic photo filter removal using convolutional neural networks. *Journal of Electronic Imaging*, 27(1):011004, 2017. 2, 5

[3] Simone Bianco, Claudio Cusano, Flavio Piccoli, and Raimondo Schettini. Learning parametric functions for color image enhancement. In *International Workshop on Computational Color Imaging*, pages 209–220. Springer, 2019. 2, 5

[4] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 97–104, 2011. 3, 4, 5, 6

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 5

[6] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):118, 2017. 2, 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 4

[8] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings*

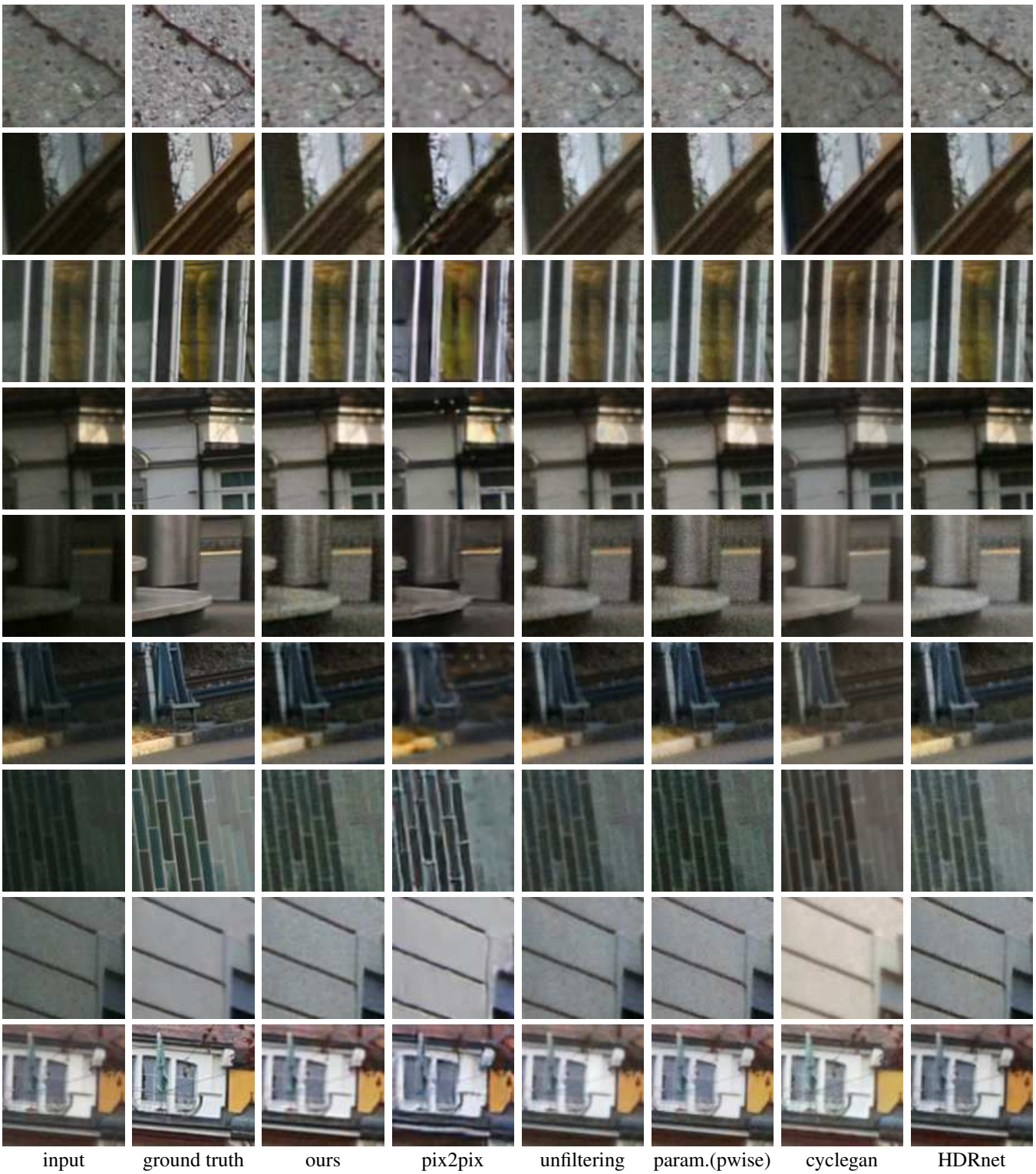


Figure 6. Visual comparison of the presented method with the ground-truth and the other methods in the state of the art applied to a sample of images from the DSLR Photo Enhancement Dataset (DPED) [8] dataset. As it's possible to see in the third column, our method is able to adjust the tone of the input picture (first column) so that is very similar to the target tone of the ground-truth (second column).

Figure 7. One example of full-resolution image processed by our method. Note that our method has been trained on small crops. The top row shows the input image and two zoomed out regions. Second row shows the output of the proposed system along with the same two regions.

of the IEEE International Conference on Computer Vision, pages 3277–3285, 2017. 2, 4, 5, 6, 7

[9] Andrey Ignatov, Radu Timofte, et al. Ntire 2019 challenge on image enhancement: Methods and results. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 3, 4, 5

[10] Andrey Ignatov, Radu Timofte, Thang Van Vu, Tung Minh Luu, Trung X Pham, Cao Van Nguyen, Yongwoo Kim, Jae-Seok Choi, Munchurl Kim, Jie Huang, et al. Pirm challenge on perceptual image enhancement on smartphones: report. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 4

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. 2, 5

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[13] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 4

[14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[15] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):11, 2016. 2

[16] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015. 2

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 5